

Sobre los dilemas éticos de la evaluación de programas sociales y su normatividad

On the ethical dilemmas of evaluating social programs and regulations

Curtis Huffman Espinosa*

Resumen

Basar la asignación de presupuesto en los resultados de las acciones públicas es una tendencia relativamente nueva en nuestro país y busca asentar la administración de los recursos públicos en la evidencia más científica posible de la eficacia y eficiencia de dichas acciones. El presente artículo trata de ajustar el fiel de la balanza que sopesa los riesgos y beneficios esperados de la evaluación de programas y políticas de desarrollo social poniendo en perspectiva el desarrollo de experimentos sociales, desde un punto de vista epistemológico. Nuestro argumento consiste en mostrar que la creencia dogmática de que las evaluaciones de impacto experimentales gozan de un estatus epistemológico privilegiado puede conducir a la modificación injustificada de la estructura de riesgos que enfrentan los sujetos de desarrollo en detrimento de estos últimos.

Palabras clave:

- Evaluación de programas
- Experimentos sociales
- Dilemas éticos
- Normatividad

Abstract

Evidence based budgeting is a relatively new movement in Mexico that seeks to put public management on a firmer scientific footing. This paper strives to adjust the pointer of the balance that weighs the expected risks and benefits of federal social program evaluation by examining the epistemological grounds of social experiments. Our argument consists in showing that the dogmatic believe that social experiments carry special scientific weight may lead to an unjustified modification of the expected risk and benefits structure that face the subjects of development against their interests. With this paper we hope to start a public discussion of the ethical matters of social program evaluation and the dearly need of an institutional framework that safeguards the social rights of the people.

Keywords:

- Program Evaluation
- Social Experiments
- Ethical Dilemmas
- Regulation

JEL: C82, C93, D63

Introducción

La evaluación de los programas y acciones públicas orientadas al desarrollo social, como toda investigación científica aplicada al estudio del ser humano, expone necesariamente a los sujetos de estudio a algún tipo de riesgo. Desde la simple recolección de datos hasta la asignación de tratamientos, la evaluación de programas de desarrollo social modifica la estructura de riesgos y beneficios esperados de los sujetos de desarrollo. Esta intervención que resulta de cualquier ejercicio de evaluación, si bien algunas intervenciones son más

* Es Maestro en Economía por El Colegio de México, donde actualmente estudia el doctorado en economía.. ■ ■ ■

“invasivas” que otras, no puede más que justificarse exclusivamente en la proporción de los beneficios esperados de cada investigación particular. De ahí que resulte de la mayor importancia tener una idea clara –epistémico–crítica– de los beneficios esperados de este tipo de investigación científica aplicada.

No es una exageración afirmar que en la práctica de la evaluación de programas se considera que evaluaciones experimentales¹ –un tipo de evaluación particularmente invasivo que generalmente proyecta retener o aplazar el apoyo de los programas y acciones de desarrollo públicas– cuentan con virtudes epistémicas especiales que las posicionan como la evidencia más objetiva y científica posible –a menudo se considera, incluso, que éstas son esenciales para probar (o extraer cualquier conclusión) de manera verdaderamente científica del impacto o efectividad de un programa de desarrollo.

El objetivo de este breve artículo es poner en perspectiva el desarrollo de evaluaciones experimentales a los programas de desarrollo social, desde un punto de vista epistemológico, para mostrar que la creencia dogmática de que las evaluaciones de impacto experimentales garantizan la validez interna de las mismas puede conducir a la modificación injustificada de la estructura de riesgos que enfrentan los sujetos de desarrollo en detrimento de sus derechos sociales.

Nuestra posición, que habremos de desarrollar en las siguientes secciones, puede resumirse como sigue. La cuestión no es oponerse o respaldar los experimentos sociales irreflexivamente a manera de principio, sino decidir cuál marco de evaluación es el más apropiado dependiendo del contexto de cada programa y acción de desarrollo atendiendo siempre a los problemas éticos inherentes a la investigación social aplicada. La posibilidad y, consecuentemente, la decisión de aleatorizar en una evaluación determinada depende de sopesar críticamente el valor de la información que se espera obtener y el efecto que este tipo de evaluaciones tienen sobre la estructura de riesgos y beneficios que enfrentan los sujetos de desarrollo.

En la siguiente sección se revisan los argumentos detrás de considerar al experimento aleatorio como condición *sine qua non* para la obtención de evidencia del impacto de un programa o acción de desarrollo social. En la tercera sección se analizan los supuestos detrás de estos argumentos y por último se resalta la importancia de mantener una posición epistemológica crítica de los aspectos metodológicos de la evaluación de los programas sociales para sopesar los riesgos y beneficios de las evaluaciones experimentales.

¹ También referidas como experimentos aleatorios, experimentos sociales, evaluaciones de asignación aleatoria o RCT por sus siglas en inglés (Randomized Controlled Trials).

a distribuirse de manera aproximadamente igual en ambos. Además, una vez conformados los dos grupos es nuevamente el azar quién decide cuál será el experimental y cuál el de control” (Cortés, 2008, p. 76).

De acuerdo con esta influyente línea de argumentación, la aleatorización puede, de un solo golpe, generar grupos estadísticamente indistinguibles no sólo en términos de todos los posibles factores confusores –los factores explicativos distintos de **X** cuyos efectos podrían confundirse con los de la variable experimental– observables y conocidos, sino también de los no observables e incluso insospechados. Esto es, la aleatorización garantiza la validez interna del experimento; es decir, que el grupo de comparación sea válido, es decir, que produzca una estimación válida de lo que habría ocurrido si el tratamiento no hubiera tenido lugar.

Siguiendo en argumento desplegado por Cortés (2008):

“En la medida que **la aleatoriedad haga bien su trabajo** se tendrá control sobre el efecto de las variables confusoras por lo que, en principio, se podrá sostener que la diferencia que se registre en la variable dependiente después de la operación de la variable experimental será consecuencia de ella y no de las variables confusoras, o más precisamente que la diferencia que se observa en **Y** entre ambos grupos al comienzo del experimento es distinta a la que se registra después que se manipuló la variable explicativa. Es claro que si **la aleatorización fue exitosa** las diferencias que se observen en las mediciones *ex ante* sólo serán explicadas por fluctuaciones de azar...

Este diseño [el experimental], además de controlar las variables que diferencian a los individuos, permite hacer lo propio con los impactos de las variables externas que llevarían erróneamente a adjudicar el efecto causal a la variable experimental garantizando así la validez interna.” (Cortés, 2008, p. 77, énfasis añadido).

O bien, que, una vez que se lleva a cabo el experimento hay buenas razones para argumentar que se han controlado todas las fuentes que atentan contra la validez interna de la evaluación; esto es, que se cuenta con un grupo de comparación que produce una estimación válida de lo que hubiera ocurrido si el programa no hubiera tenido lugar.

cuenta la posibilidad de la existencia de un número indefinidamente grande de posibles variables confusoras.

Incluso si hay una pequeña probabilidad de que un factor confusor esté desbalanceado, dado que hay k posibles de factores confusores, entonces, parecería seguirse que la probabilidad de que alguno de estos factores esté desbalanceado podría ser, hasta donde sabemos, muy alta. Pensando en posibles variables confusoras independientes, la probabilidad de que al menos una de ellas muestre una diferencia significativa entre los grupos de tratamiento y control, a un nivel de significancia α , que para un número modesto de 10 covariables y un nivel de significancia de 5%, esta probabilidad es igual a 40% (Morgan y Rubin, 2012).

Ciertamente, la sugerencia de los más acérrimos defensores de las evaluaciones experimentales es que siempre que se encuentren disponibles datos anteriores a la exposición a la intervención, estas eventualidades pueden mitigarse revisando el balance de posibles variables confusoras antes de que el experimento tenga lugar.

Así, por ejemplo, Cortés (2008) afirma sobre la evaluación del programa Progreso Oportunidades que:

“En efecto, la aleatorización del modelo experimental genera dos grupos equivalentes de acuerdo con las leyes de la estadística, pero en el caso en que la selección es **parcialmente gobernada por el azar** es necesario estudiar si al inicio de la aplicación del Programa los grupos no presentaban diferencias significativas en las variables que podrían introducir sesgos en la identificación de los efectos inducidos por él, es decir, que no difieren en otras variables que tienen relación con Y.” (Cortés, 2008, p. 80).

Una vez que se ha aceptado que en cualquier asignación aleatoria existen factores observables que pueden estar desbalanceados en ambos grupos –y los más acérrimos defensores de los experimentos aleatorios aceptan esto; aunque curiosamente sugieran re-aleatorizar (o emplear alguna otra técnica estadística apropiada para controlar por dicha variable) hasta encontrar balance antes de deliberadamente balancearles (Morgan y Rubin, 2012)–, entonces parece difícil de negar que es mejor contar con un grupo de control y experimental deliberadamente pareado, en cuanto a evitar que posibles confusores observables contaminen las observaciones en los resultados de interés, que dejarlo a la casualidad de una moneda.

hacia los resultados de otro tipo evaluaciones menos invasivas –que modifican en menor medida la estructura de riesgos y beneficios que enfrentan los individuos– que, cuando son diseñadas y conducidas cuidadosamente, pueden aportar la evidencia científica necesaria para la toma de decisiones.³ Nadie podría estar en contra de la aleatorización en todos los casos. Claramente la idea de que se debería aleatorizar siempre que se pueda está motivada por el deseo de ser tan científico como sea posible y, consecuentemente, tomar control de la evaluación esencialmente con el objetivo de intentar descartar otras explicaciones alternativas a la de que los resultados observados son consecuencia del programa bajo escrutinio. Y todo mundo está de acuerdo, en principio, en que la toma de decisiones en política pública tiene al menos que considerar, entre otras cosas, la mejor evidencia científica posible.

Mas la mejor evidencia científica posible se obtiene cuando se ha logrado eliminar las explicaciones alternativas plausibles a cualquier diferencia observada en las variables de resultado entre el grupo experimental y el grupo control. Esto significa controlar por todas las alternativas plausibles; mas, como hemos argumentado, esta validez interna no es equivalente a llevar a cabo estudios con control aleatorio. Toda vez que son el conocimiento teórico y la evidencia científica acumulada los que indican cuáles de estos factores alternativos son más plausibles de ser confundidos con los efectos de la intervención bajo escrutinio, la inferencia causal es eminentemente cualitativa.

¿A quién le interesan estos problemas epistemológicos de la evaluación de impacto?

Nos hemos embarcado en esta revisión crítica de los supuestos detrás de los estudios con control aleatorio con la intención de examinar críticamente qué cuenta como evidencia creíble en la investigación científica aplicada a la evaluación. Esto es relevante no sólo por el deseo de contar con la mejor evidencia científica posible sino por el pleno reconocimiento de que toda evaluación puede tener efectos no esperados sobre los sujetos objetos de evaluación; y que algunos de ellos pueden ser negativos, sobre todo cuando se proyecta retener o aplazar el apoyo de los programas en el curso de la evaluación.

Es así que detrás de toda evaluación hay un tema ético. Como queda claro en los argumentos desplegados por Cortés (2008, p. 80), los asuntos

³ El otro lado de la moneda de considerar a las evaluaciones experimentales como “el estándar de oro” es la degradación de estudios observacionales.

efectiva y si la evaluación de la política de desarrollo social descansa sobre una base ética sólida.

Lo anterior obliga a una amplia reflexión que involucre a operadores de programas, profesionales de la evaluación, al Consejo Nacional de Evaluación de la Política de Desarrollo Social –en su calidad de institución encargada de coordinar y normar la evaluación de los programas y políticas de desarrollo social–, la sociedad en general, así como a expertos en áreas como la ética de la investigación y los Derechos Humanos con el propósito de promover un debate acerca de las implicaciones éticas asociadas a la evaluación de los programas.

Bibliografía

- Campbell, D. T. (1957), Factors relevant to the validity of experiments in social settings, *Psychological bulletin*, 54(4), 297.
- Cortés, F., Latapí, A. E., & de la Rocha, M. G. (2008), *Metodo científico y política social: a propósito de las evaluaciones cualitativas de los programas sociales*, Colegio de Mexico AC.
- Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2011), *La evaluación de impacto en la práctica*, Banco Mundial.
- Morgan, K. L., & Rubin, D. B. (2012), Rerandomization to improve covariate balance in experiments, *The Annals of Statistics*, 40(2), 1263–1282.
- Orozco, M., Parker, S., & Hernández, D. (2000), El modelo de evaluación de Progres. Secretaría de Desarrollo Social. Más oportunidades para las familias pobres: evaluación de resultados del programa de educación, salud y alimentación, *Metodología de la evaluación de Progres*, 1–29.
- Shadish, W. R., & Cook, T. D. (1999), Comment–design rules: More steps toward a complete theory of quasi-experimentation, *Statistical Science*, 14(3), 294–300.
- Stanley, J. C., & Campbell T, D. (2012), *Diseños experimentales y cuasiexperimentales en la investigación social*. Amorrortu.
- Worrall, J. (2002), What evidence in evidence-based medicine?, *Philosophy of Science*, 69(S3), S316–S330.
- Worrall, J. (2007), Why there’s no cause to randomize, *The British Journal for the Philosophy of Science*, 58(3), 451–488.